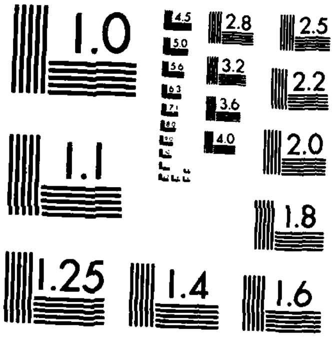
AD-A125 788 LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS BY ORDINARY LEAST SQUARES(U) RAND CORP SANTA MONICA CA G W HAGGSTROM MAR 82 RAND/P-6811 1/1

UNCLASSIFIED F/G 12/1 NL

END
FILMED
DTIC


1.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

2

LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS BY ORDINARY LEAST SQUARES

Gus W. Haggstrom

March 1982

P-6811

The Rand Paper Series

The Rand Corporation
Santa Monica, California 90406

LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS BY ORDINARY LEAST SQUARES

Gus W. Haggstrom

March 1982

This paper is based on research supported by Rand corporate funds and a grant from the Department of Health and Human Services. Views and opinions expressed in the paper are the author's own and do not necessarily reflect those of Rand or its research sponsors. The author wishes to thank William Lisowski of American Telephone and Telegraph Company and Dan Relles and John Rolph of Rand for their helpful suggestions.

LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS BY ORDINARY LEAST SQUARES

Gus W. Haggstrom

If the observations for fitting a polytomous logistic regression model satisfy certain normality assumptions, the maximum likelihood estimates of the regression coefficients are the discriminant function estimates. This paper shows that these estimates, their unbiased counterparts, and associated test statistics for variable selection can be calculated using ordinary least squares regression techniques, thereby providing a convenient procedure for performing discriminant analysis and fitting logistic regression models in the normal case. If the normality assumptions are violated, the discriminant function estimates and test statistics afford readily calculated alternatives to other procedures for fitting logistic regression models, such as the conditional maximum likelihood estimates, that present theoretical and computational difficulties. Empirical evidence is provided to show that the results of fitting logistic regression models using the discriminant function approach often agree closely with those obtained by conditional maximum likelihood.

1. INTRODUCTION

R. A. Fisher (1936) provided a convenient mnemonic derivation of the linear discriminant function based on samples from two multivariate normal distributions. He showed that a multiple of the discriminant function coefficient vector could be obtained by fitting a linear equation by least squares using the components of the observation vectors as independent variables and a dichotomous dependent variable to separate the individuals in the two samples. Later it was shown that the t- and F-statistics associated with this least squares procedure provide valid tests of hypotheses pertaining to the discriminant coefficients. This paper extends these results to the case of three or more populations and shows how the analogous logistic regression model in the normal case can be fitted and tested using least squares techniques.

The logistic regression model arises in quantifying the dependence of a polytomous (categorical) variable y on a q-dimensional vector $\underset{\sim}{x}$ of explanatory variables. Logistic regression (or "logit analysis") is related to discriminant analysis in that the variable y may reflect membership in one of several populations, in each of which the vector $\underset{\sim}{x}$ has a multivariate normal distribution.

To explore this relationship, we first consider the general classification problem. Suppose that an individual is drawn at random from a population consisting of m disjoint subpopulations π_1, π_2, ..., π_m, and consider the problem of classifying the individual into one of the subpopulations on the basis of a q-dimensional vector $\underset{\sim}{x}$ of measurements on this individual. Let y be the random variable having the value j for individuals in π_j, and let $p_j = P(y = j)$ denote the prior probability of drawing an individual from π_j.

If the conditional density of $\underset{\sim}{x}$ in π_j with respect to some measure μ on R^q is $f_j(\underset{\sim}{x})$, the posterior probability that the individual belongs to π_j given $\underset{\sim}{x}$ is

$$p(j|\underset{\sim}{x}) = P(y = j|\underset{\sim}{x}) = p_j f_j(\underset{\sim}{x})/\Sigma_{k=1}^m p_k f_k(\underset{\sim}{x}) \quad . \qquad (1.1)$$

Among rules for classifying an individual into one of the subpopulations π_j on the basis of $\underset{\sim}{x}$, the rule that decides y = i when $p(i|\underset{\sim}{x}) = \max_j p(j|\underset{\sim}{x})$ maximizes the probability of correct classification (Ferguson, 1967, p. 292). In practice, the density functions $f_j(\underset{\sim}{x})$ and the prior probabilities p_j are usually unknown, and statistical models are often posited in which the conditional probabilities are expressed as simple functions of parameters that can be estimated from a training set of n observations $(\underset{\sim}{x}_i, y_i)$, i = 1, 2, ..., n.

A logistic regression model for the pair $(\underset{\sim}{x}, y)$ is characterized by the condition that the probabilities $p(j|\underset{\sim}{x})$ are expressible in the form

$$p(j|\underset{\sim}{x}) = \exp(\gamma_j + \underset{\sim}{\delta}_j{}'\underset{\sim}{x})/\Sigma_{k=1}^m \exp(\gamma_k + \underset{\sim}{\delta}_k{}'\underset{\sim}{x}) \qquad (1.2)$$

for some set of parameters γ_j and $\underset{\sim}{\delta}_j = (\delta_{j1}, \ldots, \delta_{jq})'$. In the dichotomous case (m = 2), this reduces to the binary logistic form

$$p(1|\utilde{x}) = 1/[1 + e^{-(\alpha + \utilde{\beta}'\utilde{x})}] \; , \tag{1.3}$$

if one sets $\alpha = \gamma_1 - \gamma_2$ and $\utilde{\beta} = \utilde{\delta}_1 - \utilde{\delta}_2$.

The parameters γ_j and $\utilde{\delta}_j$ in (1.2) are not uniquely determined, because the probabilities $p(j|\utilde{x})$ remain unchanged if one multiplies the numerator and denominator in (1.2) by $\exp(a + \utilde{b}'\utilde{x})$ for any a and $\utilde{b}$. One way to specify the parameters uniquely is to incorporate side conditions, such as $\Sigma\, \gamma_k = 0$ and $\Sigma\, \utilde{\delta}_k = \utilde{0}$, or $\gamma_m = 0$ and $\utilde{\delta}_m = \utilde{0}$. Alternatively, one can specify these parameters as functions of other parameters that index the joint distribution of $\utilde{x}$ and y. The latter method will be used in treating the normal case below.

It follows from (1.1) that the pair $(\utilde{x}, y)$ satisfies a logistic regression model whenever $\log[f_j(\utilde{x})/f_m(\utilde{x})]$ is a linear function in $\utilde{x}$ for $j = 1, 2, \ldots, m - 1$. In particular, this holds if the conditional densities belong to an exponential family

$$f_j(\utilde{x}) = C(\utilde{\theta}_j)\; h(\utilde{x})\; \exp(\utilde{\theta}_j{}'\utilde{x}) \tag{1.4}$$

where $\utilde{\theta}_j$ is a q-dimensional vector of parameters. Day and Kerridge (1967) provided a slightly different specification by adopting densities of the form

$$f_j(\utilde{x}) = c_j \; \exp[-(\utilde{x} - \utilde{\mu}_j)'\utilde{\Sigma}^{-1}(\utilde{x} - \utilde{\mu}_j)/2]\; \varphi(\utilde{x}) \tag{1.5}$$

for some q-dimensional vector $\utilde{\mu}_j$ and some nonsingular q×q covariance matrix $\utilde{\Sigma}$. This can be written in the form (1.4) with $\utilde{\theta}_j = \utilde{\Sigma}^{-1}\utilde{\mu}_j$.

The normal case that gives rise to a logistic regression model is the case in which $\utilde{x}$ has a multivariate normal distribution $N_q(\utilde{\mu}_j, \utilde{\Sigma}_j)$

and the covariance matrices satisfy $\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$. Here, the density of $\boldsymbol{x}$ in π_j is

$$f_j(\boldsymbol{x}) = (2\pi)^{-q/2}|\boldsymbol{\Sigma}|^{-1/2} \exp[-(\boldsymbol{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)/2] \ . \tag{1.6}$$

It follows from (1.1) that the conditional probabilities $p(j|\boldsymbol{x})$ satisfy a logistic regression model (1.2) with the parameters equal to the discriminant function coefficients

$$\begin{aligned} \boldsymbol{\delta}_j &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j \ , \\ \gamma_j &= \log p_j - \boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j/2 \ . \end{aligned} \tag{1.7}$$

In the dichotomous case, the parameters of the binary logistic model (1.3) are given by

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ \alpha &= \log(p_1/p_2) - \boldsymbol{\beta}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 \ . \end{aligned} \tag{1.8}$$

In treating the estimation of the parameters in (1.7) and (1.8), we assume there is a training set of n independent observations $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, \dots, n$, such that for any pair $(\boldsymbol{x}, y)$ the distribution of $\boldsymbol{x}$ given $y = j$ is $N_q(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Let n_j be the number of observations for which $y_i = j$. Two cases will be considered:

Case I: The variables y_i are random with $P(y_i = j) = p_j$.

Case II: The variables y_i are constants, i.e., the observations arise from separate samples of fixed sizes $n_1, \dots, n_m$ from populations $\pi_1, \dots, \pi_m$.

In Case I, the maximum likelihood estimators (MLEs) of the parameters p_j, $\boldsymbol{\mu}_j$, and $\boldsymbol{\Sigma}$ are the values that maximize the likelihood function

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} [p_j f_j(\boldsymbol{x}_i)]^{v_{ji}} = \prod_{j=1}^{m} p_j^{n_j} \prod_{i=1}^{n} [f_j(\boldsymbol{x}_i)]^{v_{ji}} \ , \tag{1.9}$$

where $v_{ji} = 1$ if $y_i = j$ and $v_{ji} = 0$ otherwise. The likelihood function for Case II is the same except that the factors involving p_j are missing. In either case, the MLEs of the parameters γ_j and $\underset{\sim}{\delta}_j$ are the discriminant function estimators obtained by substituting the MLEs $\hat{\underset{\sim}{\mu}}_j$ and $\hat{\underset{\sim}{\Sigma}}$ in (1.7). By well-known results for Case II (e.g., Anderson, 1958, p. 248), the MLE of $\underset{\sim}{\mu}_j$ is the sample mean vector

$$\hat{\underset{\sim}{\mu}}_j = \bar{\underset{\sim}{x}}_j = \sum_i v_{ji}\underset{\sim}{x}_i/n_j \ , \qquad (1.10)$$

and the MLE of $\underset{\sim}{\Sigma}$ is $\hat{\underset{\sim}{\Sigma}} = \underset{\sim}{A}/n$, where $\underset{\sim}{A}$ is the pooled sum of squares and cross products matrix

$$\underset{\sim}{A} = \sum_{i=1}^{n} \sum_{j=1}^{m} v_{ji}(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_j)(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}_j)' . \qquad (1.11)$$

If the values of p_j are unknown in Case I, the MLE of p_j is $\hat{p}_j = n_j/n$. In Case II, we shall assume the p_j's are known.

Thus, the MLEs of the parameters in (1.7) are

$$\begin{aligned} \hat{\underset{\sim}{\delta}}_j &= \hat{\underset{\sim}{\Sigma}}^{-1}\bar{\underset{\sim}{x}}_j \\ \hat{\gamma}_j &= \log \hat{p}_j - \bar{\underset{\sim}{x}}_j'\hat{\underset{\sim}{\Sigma}}^{-1}\bar{\underset{\sim}{x}}_j/2 . \end{aligned} \qquad (1.12)$$

In the dichotomous case (1.8), the MLEs are

$$\begin{aligned} \hat{\underset{\sim}{\beta}} &= \hat{\underset{\sim}{\Sigma}}^{-1}(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) \\ \hat{\alpha} &= \log (\hat{p}_1/\hat{p}_2) - \hat{\underset{\sim}{\beta}}'(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)/2 . \end{aligned} \qquad (1.13)$$

While a number of statistical packages exist for calculating the discriminant function estimates directly, few provide test statistics for performing variable selection, and they often lack the versatility for making transformations, deleting variables (or cases), plotting, and treating missing values that linear model practitioners are accustomed to. We now show how these estimates, their unbiased counterparts, and associated test statistics can be calculated by applying least squares procedures to linear models.

2. THE DICHOTOMOUS CASE

For the case $m = 2$, we redefine the variables y_i to have values 1 and 0, instead of 1 and 2, and let $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ denote the vector of dummy variables indicating membership in π_1. Let a and $\mathbf{b}$ denote the "intermediate least squares" (ILS) estimates that result from treating the observations $(\mathbf{x}_i, y_i)$ as if they satisfied a linear model

$$y_i = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + e_i \ , \tag{2.1}$$

and let SS_e denote the residual sum of squares from this regression:

$$SS_e = \sum_{i=1}^{n} (y_i - a - \mathbf{b}'\mathbf{x}_i)^2 \ . \tag{2.2}$$

<u>Theorem 1</u>. The MLEs of the logistic regression coefficients in (1.8) are related to the ILS estimates a and $\mathbf{b}$ by

$$\begin{gathered} \hat{\boldsymbol{\beta}} = K\mathbf{b}, \\ \hat{\alpha} = \log(\hat{p}_1/\hat{p}_2) + K(a - 1/2) + n(n_1^{-1} - n_2^{-1})/2 \ , \end{gathered} \tag{2.3}$$

where $K = n/SS_e$.

Before proceeding with the proof, we first note that, since $\hat{p}_1/\hat{p}_2 = n_1/n_2 = \bar{y}(1 - \bar{y})$ in Case I, these estimates can be readily calculated by hand from just the values of a, $\mathbf{b}$, n, SS_e, and $\bar{y}$. Also, as will be seen below, the standard errors and t-statistics for the logistic regression coefficients $\hat{\beta}_k$ are readily obtained from those associated with the ILS estimates.

In Fisher's original mnemonic procedure for deriving an unspecified multiple of the discriminant function coefficients, he used a dichotomous dependent variable $\underset{\sim}{u}$ having the values n_2/n and $- n_1/n$ instead of 1 and 0. Since the values u_i are the centered values $y_i - \bar{y}$, the values of $\underset{\sim}{b}$ and SS_e are the same for both choices. Previous writers (Warner, 1961; Lachenbruch, 1975, p. 26) have derived formulas for the factor K in (2.3) that are not as readily calculated, given the value of SS_e.

To prove Theorem 1, we follow Fisher (1938) and observe that $\underset{\sim}{b}$ satisfies the normal equations

$$\underset{\sim}{Z}'\underset{\sim}{Z}\underset{\sim}{b} = \underset{\sim}{Z}'\underset{\sim}{u} \qquad (2.4)$$

where $\underset{\sim}{Z}$ is the nxq matrix of centered values of the x_{ij}'s. These equations can be rewritten in the form

$$(\underset{\sim}{A} + c\underset{\sim}{d}\underset{\sim}{d}')\underset{\sim}{b} = c\underset{\sim}{d} \ , \qquad (2.5)$$

where $\underset{\sim}{d} = \bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2$, $c = n_1n_2/n$, and $\underset{\sim}{A} = n\hat{\underset{\sim}{\Sigma}}$ is the pooled sum of squares and cross products matrix. Thus, $\underset{\sim}{A}\underset{\sim}{b} = c(1 - \underset{\sim}{b}'\underset{\sim}{d})\underset{\sim}{d}$, implying that

$$\underset{\sim}{b} = c(1 - \underset{\sim}{b}'\underset{\sim}{d})\underset{\sim}{A}^{-1}\underset{\sim}{d} = \hat{\underset{\sim}{\beta}}/K \qquad (2.6)$$

where

$$K = n/c(1 - \underset{\sim}{b}'\underset{\sim}{d}) \ . \qquad (2.7)$$

Since the sum of squares due to regression is

$$SS(reg) = \underset{\sim}{b}'\underset{\sim}{Z}'\underset{\sim}{u} = c\underset{\sim}{b}'\underset{\sim}{d} \qquad (2.8)$$

and the total sum of squares about the mean is

$$SS(tot) = n_1(1 - n_1/n)^2 + n_2(-n_1/n)^2 = c \ , \qquad (2.9)$$

the residual sum of squares is

$$SS_e = c(1 - \underset{\sim}{b}'\underset{\sim}{d}) = n/K \ . \qquad (2.10)$$

Hence, from (2.6) and (2.10), $\hat{\beta} = Kb$ where $K = n/SS_e$.

To show that $\hat{\alpha}$ in (1.12) can be written in the form (2.2), it suffices to show that

$$-\underset{\sim}{\hat{\beta}}'(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2)/2 = K(a - 1/2) + n(n_1^{-1} - n_2^{-1})/2 \qquad (2.11)$$

or

$$\underset{\sim}{b}'(\bar{x}_1 + \bar{x}_2) = -2a + 1 - SS_e(n_1^{-1} - n_2^{-1}) \; . \qquad (2.12)$$

Let $\underset{\sim}{x}_{jk}$, $k = 1, \ldots, n_j$, denote the $\underset{\sim}{x}_i$ vectors for the n_j observations from π_j, and let $\hat{y}_{jk}$ denote the fitted value corresponding to $\underset{\sim}{x}_{jk}$. Then

$$\underset{\sim}{b}'\bar{\underset{\sim}{x}}_j = \Sigma\, \underset{\sim}{b}'\underset{\sim}{x}_{jk}/n_j = \Sigma\, (\hat{y}_{jk} - a)/n_j = -a + \Sigma\, \hat{y}_{jk}/n_j \; . \qquad (2.13)$$

Also,

$$\Sigma\, \hat{y}_{1k} = \hat{\underset{\sim}{y}}'\underset{\sim}{y} = \underset{\sim}{y}'\underset{\sim}{y} - (\underset{\sim}{y} - \hat{\underset{\sim}{y}})'\underset{\sim}{y} = n_1 - (\underset{\sim}{y} - \hat{\underset{\sim}{y}})'(\underset{\sim}{y} - \hat{\underset{\sim}{y}}) = n_1 - SS_e \qquad (2.14)$$

and

$$\Sigma\, \hat{y}_{2k} = \Sigma\, y_i - \Sigma\, \hat{y}_{1k} = n_1 - (n_1 - SS_e) = SS_e. \qquad (2.15)$$

The result follows from substituting (2.13)-(2.15) in the left member of (2.12), completing the proof of Theorem 1.

It is well-known that the F-statistic for testing the hypothesis H: $\underset{\sim}{\beta} = 0$ (or, equivalently, $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$), calculated as though the linear model (2.1) applied with normally distributed errors $e_i \sim N(0, \sigma^2)$, is a multiple of Hotelling's two-sample T^2 statistic

$$T^2 = c\underset{\sim}{d}'\underset{\sim}{S}^{-1}\underset{\sim}{d} = cD_q^{\,2} \; , \qquad (2.16)$$

where $\underset{\sim}{S} = \underset{\sim}{A}/(n - 2)$ is the usual unbiased estimator of $\underset{\sim}{\Sigma}$, and $D_q^{\,2}$ is Mahalanobis' D^2 statistic (Fisher, 1938). From (2.6)-(2.8), we see that

$$SS(reg) = c\underset{\sim}{b}'\underset{\sim}{d} = cn\underset{\sim}{d}'\underset{\sim}{A}^{-1}\underset{\sim}{d}/K = cD_q^{\,2}(SS_e)/(n - 2). \qquad (2.17)$$

Hence, the F-statistic is

$$F = (n - q - 1)SS(reg)/q(SS_e) = (n - q - 1)T^2/q(n - 2). \qquad (2.18)$$

Under the assumption that the observation vectors $\underset{\sim}{x}_i$ are sampled from two multivariate normal distributions $N_q(\underset{\sim}{\mu}_j, \underset{\sim}{\Sigma})$, $j = 1, 2$, it follows that $F \sim F(q, n - q - 1)$ under H (Anderson, 1958, p. 109). Lehmann (1959) showed that this test is the uniformly most powerful (UMP) invariant test of H.

To test H_1: $\beta_{p+1} = \ldots = \beta_q = 0$, one can follow the linear model paradigm to calculate an F-statistic F_1 comparing SS_e with the residual sum of squares SS_ω calculated after omitting the last $q - p$ components of $\underset{\sim}{x}$ as independent variables. Since it follows from (2.17) that

$$SS_e = C/(1 + CD_q^2) \qquad (2.19)$$

where $C = c/(n - 2)$, we see that

$$F_1 = k(SS_\omega/SS_e - 1) = kC(D_q^2 - D_p^2)/(1 + CD_p^2) \qquad (2.20)$$

where $k = (n - q - 1)/(q - p)$. Rao (1946, 1948) derived F_1 as the likelihood ratio test statistic for testing H_1 in the two-sample multivariate normal case and showed that $F_1 \sim F(q - p, n - q - 1)$ under H_1. Giri (1964) proved that Rao's test is the UMP invariant similar test of H_1. These results are summarized in the following theorem.

Theorem 2. The F-statistics (2.18) and (2.20) derived from the linear model paradigm provide valid tests of H: $\underset{\sim}{\beta} = 0$ and H_1: $\beta_{p+1} = \ldots = \beta_q = 0$.

To extend the linear model paradigm further, we define the standard error of $\hat{\beta}_k$ and the t-statistic t_k for testing H_2: $\beta_k = 0$ by

$$\text{s.e.}(\hat{\beta}_k) = K[\text{s.e.}(b_k)] \ ,$$
$$t_k = \hat{\beta}_k/\text{s.e.}(\hat{\beta}_k) = b_k/\text{s.e.}(b_k) \ , \qquad (2.21)$$

where s.e.(b_k) is the standard error of b_k calculated from fitting the y_i's to the $\underset{\sim}{x}_i$'s by OLS. Then it follows from specializing Rao's result to the case $p = q - 1$ that t_k provides a valid test of H_2 in the sense that $t_k^2 \sim F(1, n - q - 1)$. This suggests, but does not prove, that t_k has a t distribution under H_2. While this result will be proved below, it is not true that $(\hat{\beta}_k - \beta_k)/\text{s.e.}(\hat{\beta}_k)$ has a t distribution when $\beta_k \neq 0$. The problems of providing unbiased estimators and confidence intervals for β_k will be treated later.

3. THE POLYTOMOUS CASE

The development above for the dichotomous case provides little insight as to why following the linear model paradigm might lead to valid tests and estimates for the logistic regression model. To further illuminate the dichotomous case and provide a basis for establishing analogous results for the polytomous case, we begin by considering how the parameters of the logistic regression model are related to certain regression coefficients whose MLEs are least-squares estimators.

In the logistic regression model given by (1.2), the conditional probabilities $p(j|\underset{\sim}{x})$ are specified in terms of parameters γ_j and $\underset{\sim}{\delta}_j$ that are functions of p_j, $\underset{\sim}{\mu}_j$, and $\underset{\sim}{\Sigma}$. A second parameterization that is more convenient for this development results from dividing the numerator and denominator of (1.2) by $\exp(\gamma_m + \underset{\sim}{\delta}_m{}'x)$ and setting $\alpha_j = \gamma_j - \gamma_m$ and $\underset{\sim}{\beta}_j = \underset{\sim}{\delta}_j - \underset{\sim}{\delta}_m$. This yields the parameterization

$$p(j|\underset{\sim}{x}) = \exp(\alpha_j + \underset{\sim}{\beta}_j{}'\underset{\sim}{x})/[1 + \Sigma_{k=1}^{m-1} \exp(\alpha_k + \underset{\sim}{\beta}_k{}'\underset{\sim}{x})] \quad \text{for } j = 1, 2, \ldots, m-1; \qquad (3.1)$$

$$p(m|\underset{\sim}{x}) = 1 - \Sigma_{j=1}^{m-1} p(j|\underset{\sim}{x}).$$

It follows from (1.7) that

$$\underset{\sim}{\beta}_j = \underset{\sim}{\Sigma}^{-1}(\underset{\sim}{\mu}_j - \underset{\sim}{\mu}_m)$$
$$\alpha_j = \log(p_j/p_m) - \underset{\sim}{\beta}_j{}'(\underset{\sim}{\mu}_j + \underset{\sim}{\mu}_m)/2. \qquad (3.2)$$

Letting the (i,j) elements of $\underset{\sim}{\Sigma}$ and $\underset{\sim}{\Sigma}^{-1}$ be denoted by σ_{ij} and σ^{ij} in the sequel, we recall that, if $\underset{\sim}{x} \sim N_q(\underset{\sim}{\mu}_j, \underset{\sim}{\Sigma})$, then the conditional distribution of x_k, given the other components of $\underset{\sim}{x}$, is $N(\xi_{jk} + \underset{\sim}{\theta}_k{}'\underset{\sim}{x}, 1/\sigma^{kk})$ where $\underset{\sim}{\theta}_k$

is a q-dimensional vector with $\theta_{kk} = 0$, $\theta_{ki} = -\sigma^{ki}/\sigma^{kk}$ for $i \neq k$, and $\xi_{jk} = \mu_{jk} - \theta_k'\mu_j$. (See Anderson, 1958, pp. 28, 42.) The "constant terms" ξ_{jk}, $k = 1, 2, \ldots, q$, are related to the components of the logistic regression coefficient vectors $\underset{\sim}{\delta}_j = \underset{\sim}{\Sigma}^{-1}\underset{\sim}{\mu}_j$, since the k^{th} component of $\underset{\sim}{\delta}_j$ is

$$\delta_{jk} = \sum_i \sigma^{ki}\mu_{ji} = \sigma^{kk}(\mu_{jk} - \underset{\sim}{\theta}_k'\underset{\sim}{\mu}_j) = \sigma^{kk}\xi_{jk}. \tag{3.3}$$

Hence, the components of $\underset{\sim}{\beta}_j$ are given by

$$\beta_{jk} = \sigma^{kk}(\xi_{jk} - \xi_{mk}). \tag{3.4}$$

Let $v_1, v_2, \ldots, v_m$ denote the indicator variables for the subpopulations $\pi_1, \pi_2, \ldots, \pi_m$. Then it follows from the above that the components x_k of the observation vectors $\underset{\sim}{x}$ satisfy the linear model

$$\begin{aligned} x_k &= \Sigma_{j=1}^{m} \xi_{jk}v_j + \underset{\sim}{\theta}_k'\underset{\sim}{x} + e_k \\ &= \xi_{mk} + \underset{\sim}{\theta}_k'\underset{\sim}{x} + \Sigma_{j=1}^{m-1} (\xi_{jk} - \xi_{mk})v_j + e_k \end{aligned} \tag{3.5}$$

where $e_k \sim N(0, 1/\sigma^{kk})$. By relabeling $v_1, \ldots, v_m$ as $x_{q+1}, \ldots, x_{q+m}$, this can be rewritten in the form

$$x_k = \xi_{mk} + \underset{\sim}{\theta}_k'\underset{\sim}{x} + \Sigma_{j=1}^{m-1} \theta_{k,q+j}x_{q+j} + e_k \tag{3.6}$$

where

$$\theta_{k,q+j} = \xi_{jk} - \xi_{mk} = \beta_{jk}/\sigma^{kk}. \tag{3.7}$$

By reexpressing each of the joint densities $f_j(\underset{\sim}{x})$ in the likelihood function (1.9) as a product of the conditional density of x_k times the joint density of the other components of $\underset{\sim}{x}$, we see that the MLEs of the regression coefficients in (3.6) are the least-squares estimators, and the MLE of the conditional variance $1/\sigma^{kk}$ is the residual sum of squares $SS(x_k)$ divided by n. As is well known (e.g., Rao, 1965, p. 224), these estimators can be obtained by inverting the augmented sample covariance matrix $\underset{\sim}{S}$ with elements

$$s_{ij} = \sum_{\nu} (x_{i\nu} - \bar{x}_i)(x_{j\nu} - \bar{x}_j)/n \tag{3.8}$$

for i, j = 1, 2, ..., q + m - 1. The MLEs are given in terms of the elements s^{ij} of $\underset{\sim}{S}^{-1}$ by

$$\begin{aligned} \hat{\theta}_{kj} &= - s^{kj}/s^{kk} \quad \text{for } j \neq k, \\ \hat{\sigma}^{kk} &= s^{kk} = n/SS(x_k). \end{aligned} \tag{3.9}$$

Hence, by (3.7)

$$\hat{\beta}_{jk} = \hat{\theta}_{k,q+j} s^{kk} = - s^{k,q+j} \tag{3.10}$$

for j = 1, 2, ..., m-1. Noting that the last member of (3.10) can also be obtained by using x_{q+j} (= v_j) as the regressand, we obtain that

$$\hat{\beta}_{jk} = - s^{q+j,k} = \hat{\theta}_{q+j,k}\ s^{q+j,q+j} = b_{jk}(n/SS_j) \tag{3.11}$$

where $b_{jk} = \hat{\theta}_{q+j,k}$ is the regression coefficient on x_k when v_j is regressed linearly on $x_1, \ldots, x_q$ and the other v_k's, and SS_j is the residual sum of squares.

This development indicates that the linear model paradigm for the dichotomous case can be extended to the polytomous case. The procedure consists of first fitting the observations $(\underset{\sim}{x}_i, v_{1i}, \ldots, v_{m-1,i})$ by least squares as if they satisfied the linear model

$$v_{ji} = \alpha_j + \underset{\sim}{\beta}_j{}'\underset{\sim}{x}_i + \sum_{k\neq j} \gamma_{jk} v_{ki} + e_{ji} \tag{3.12}$$

for each j. This provides ILS estimates a_j and $\underset{\sim}{b}_j$ of α_j and $\underset{\sim}{\beta}_j$, as well as the residual sum of squares SS_j, which can then be transformed to yield the MLEs. The process can be summarized as follows:

Theorem 3. The MLEs of the polytomous logistic regression coefficients (3.2) are related to the ILS estimates a_j and $\underset{\sim}{b}_j$ by

$$\hat{\underset{\sim}{\beta}}_j = K_j \underset{\sim}{b}_j , \qquad (3.13)$$
$$\hat{\alpha}_j = \log(\hat{p}_j/\hat{p}_m) + K_j(a_j - 1/2) + n(n_j^{-1} - n_m^{-1})/2 ,$$

where $K_j = n/SS_j$.

The formula for $\hat{\underset{\sim}{\beta}}_j$ is a restatement of (3.11). The derivation of the formula for $\hat{\alpha}_j$ is similar to the proof for the dichotomous case. See (2.11) to (2.15).

As in the dichotomous case, these formulas for the MLEs apply whether (1) the individuals are sampled at random from the population consisting of m subpopulations $\pi_1, \ldots, \pi_m$, or (2) the observations arise from separate samples of fixed sizes $n_1, \ldots, n_m$ from $\pi_1, \ldots, \pi_m$. In the first case, the MLEs of the p_j's are $\hat{p}_j = n_j/n$; in the second, the p_j's are assumed to be known.

Next we consider whether the t-statistics derived from the linear model paradigm provide valid tests of the hypotheses H: $\beta_{jk} = 0$. Since $\beta_{jk} = \sigma^{kk} \theta_{k,q+j}$ by (3.7), H is equivalent to the hypothesis that $\theta_{k,q+j} = 0$ in (3.6). Under the Case II (separate sample) assumptions, the UMP unbiased test of H is based on the t-statistic

$$t = \hat{\theta}_{k,q+j}/\text{s.e.}(\hat{\theta}_{k,q+j}) , \qquad (3.14)$$

which has a $t(n-m-q+1)$ distribution under H. The analogous t-statistic when v_j is regressed linearly on $x_1, x_2, \ldots, x_q$ and the other v_k's is

$$t = b_{jk}/\text{s.e.}(b_{jk}) , \qquad (3.15)$$

where the standard error in the denominator is calculated as if (3.12) applied with the error terms satisfying the usual linear model assumptions. To see that these two t-statistics are identical, it suffices to recall that both can be calculated from the sample partial correlation coefficient $r_{jk.c}$ between v_j and x_k using the formula

$$t = r_{jk.c}[\nu/(1 - r_{jk.c}^2)]^{1/2} , \tag{3.16}$$

where $\nu = n - m - q + 1$.

In concluding that the t-statistic (3.15) has the same properties as those cited for the one in (3.14), one must recognize that these properties depend on the Case II assumptions. In Case I, the conclusions need qualification, because the n_j's are random variables that can be zero with positive probability. While these results and others to follow can be restated as conditional results given any nonzero values of the n_j's for which $n > m + q - 1$, we shall simply assume that the Case II assumptions apply with fixed nonzero sample sizes n_j. With this proviso, the validity of the t-statistic (3.15) can be stated as follows:

<u>Theorem 4</u>. The t-statistic (3.15) for testing H: $\beta_{jk} = 0$ derived from fitting (3.12) by ordinary least squares has a $t(\nu)$ distribution under H, and rejecting H for $|t| > t_{1-\alpha/2}(\nu)$ provides the UMP unbiased test of size α.

If we define the standard error of $\hat{\beta}_{jk}$ using

$$\text{s.e.}(\hat{\beta}_{jk}) = K_j[\text{s.e.}(b_{jk})] , \tag{3.17}$$

then $t = \hat{\beta}_{jk}/\text{s.e.}(\hat{\beta}_{jk})$ provides a valid t-statistic for testing whether $\beta_{jk} = 0$. However, it is not the case that $(\hat{\beta}_{jk} - \beta_{jk})/\text{s.e.}(\hat{\beta}_{jk})$ has a

Student's t distribution except when $\beta_{jk} = 0$. The problem of providing an approximate pivotal quantity for β_{jk} will be treated below after considering the bias of the MLEs.

By (1.12), alternative formulas for $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$ are given by

$$\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_m), \tag{3.18}$$
$$\hat{\alpha}_j = \log(p_j/p_m) - (Q_j - Q_m)/2 \ ,$$

where $Q_j = \bar{\mathbf{x}}_j'\hat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}}_j$. Das Gupta (1968) examined the moments and asymptotic distribution of the discriminant function coefficients β in the dichotomous case. A key result in his derivation is that, if $\mathbf{A}$ has a Wishart distribution $W_q(\boldsymbol{\Sigma}, N)$, then $E(\mathbf{A}^{-1}) = \boldsymbol{\Sigma}^{-1}/(N - q - 1)$. Applying this result to $\boldsymbol{\beta}_j$ and observing that the matrix $\mathbf{A} = n\hat{\boldsymbol{\Sigma}}$ in (1.11) has a $W_q(\boldsymbol{\Sigma}, n - m)$ distribution and is independent of the mean vectors, we see that $E(\hat{\boldsymbol{\beta}}_j) = n\boldsymbol{\beta}_j/(\nu - 2)$. Hence, an unbiased estimator of $\boldsymbol{\beta}_j$ is

$$\tilde{\boldsymbol{\beta}}_j = (\nu - 2)\hat{\boldsymbol{\beta}}_j/n = C_j\mathbf{b}_j \tag{3.19}$$

where $C_j = (n - m - q - 1)SS_j$.

To remove the bias in $\hat{\alpha}_j$, first note that

$$\begin{aligned} E(Q_j) &= E[E(\bar{\mathbf{x}}_j'\hat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}}_j \mid \bar{\mathbf{x}}_j)] = nE(\bar{\mathbf{x}}_j'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_j)/(\nu - 2) \\ &= n[(q/n_j) + \boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j]/(\nu - 2) \ . \end{aligned} \tag{3.20}$$

It follows that an unbiased estimator of α_j is

$$\tilde{\alpha}_j = \log(p_j/p_m) - [(\nu - 2)(Q_j - Q_m)/n - q(n_j^{-1} - n_m^{-1})]/2. \tag{3.21}$$

By (3.13), this can also be written in the form

$$\tilde{\alpha}_j = \log(p_j/p_m) + C_j(a_j - 1/2) + (n - m - 1)(n_j^{-1} - n_m^{-1})/2. \tag{3.22}$$

The unbiased estimators in (3.19) and (3.22) are functions of the sample mean vectors $\bar{\mathbf{x}}_j$ and the pooled sample covariance matrix $\hat{\boldsymbol{\Sigma}}$, which are sufficient statistics under the Case II assumptions. Moreover, $(\bar{\underset{\sim}{x}}_1, \ldots, \bar{\underset{\sim}{x}}_m, \hat{\underset{\sim}{\Sigma}})$ is complete, as can be shown by a proof analogous to that given in the one-sample case (Anderson, 1958, p. 117). It follows from the Lehmann-Scheffé Theorem that $\tilde{\alpha}_j$ and $\tilde{\underset{\sim}{\beta}}_j$ satisfy the following optimality property.

Theorem 5. The estimators $\tilde{\alpha}_j$ and $\tilde{\underset{\sim}{\beta}}_j$ given in (3.22) and (3.19) are the uniformly minimum variance unbiased estimators of α_j and $\underset{\sim}{\beta}_j$.

If one defines the standard error of $\tilde{\beta}_{jk}$ using

$$\text{s.e.}(\tilde{\beta}_{jk}) = C_j[\text{s.e.}(b_{jk})] \; , \tag{3.23}$$

then $t = \tilde{\beta}_{jk}/\text{s.e.}(\tilde{\beta}_{jk})$ has a $t(\nu)$ distribution when $\beta_{jk} = 0$ by Theorem 4. Although the pivotal quantity

$$t = (\tilde{\beta}_{jk} - \beta_{jk})/\text{s.e.}(\tilde{\beta}_{jk}) \tag{3.24}$$

only has a $t(\nu)$ distribution when $\beta_{jk} = 0$, it can still be used as an approximate pivotal quantity for generating confidence intervals. This quantity is closely related to a bona fide pivotal quantity having a $t(\nu)$ distribution suggested by the model (3.6), namely,

$$t = (\hat{\theta}_{k,q+j} - \theta_{k,q+j})/\text{s.e.}(\hat{\theta}_{k,q+j}) \; . \tag{3.25}$$

By (3.7), (3.10), (3.14), and (3.15), this can be rewritten in the form

$$\begin{aligned} t &= (\hat{\beta}_{jk}/s^{kk} - \beta_{jk}/\sigma^{kk})/(K_j/s^{kk})\text{s.e.}(b_{jk}) \\ &= (G\tilde{\beta}_{jk} - \beta_{jk})/G[\text{s.e.}(\tilde{\beta}_{jk})] \; , \end{aligned} \tag{3.26}$$

where $G = n\sigma^{kk}/\nu s^{kk} = SS(x_k)/\nu(1/\sigma^{kk})$. Noting that $SS(x_k)/\nu$ is the usual unbiased estimator of $1/\sigma^{kk}$,(the conditional variance of x_k given the other components of $\underset{\sim}{x}$), we see that $E(G) = 1$ and $\text{Var}(G) = 2/\nu$. Hence, for moderately large values of ν, omitting the factors G in (3.26) and using (3.24) instead should provide good approximations.

4. MATRIX FORMULATION

Let the augmented sample covariance matrix $\mathbf{S} = (s_{ij})$ defined in (3.8) be partitioned into

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} , \tag{4.1}$$

where $\mathbf{S}_{11}$ is the sample covariance matrix of $x_1, \ldots, x_q$, and $\mathbf{S}_{22}$ is the sample covariance matrix of $v_1, \ldots, v_{m-1}$. Then

$$\mathbf{S}^{-1} = \begin{pmatrix} \mathbf{S}_{11.2}^{-1} & -\mathbf{S}_{11.2}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} \\ -\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11.2}^{-1} & \mathbf{S}^{22} \end{pmatrix} , \tag{4.2}$$

where

$$\begin{aligned} \mathbf{S}_{11.2} &= \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} , \\ \mathbf{S}^{22} &= \mathbf{S}_{22}^{-1} + \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11.2}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1} . \end{aligned} \tag{4.3}$$

The submatrices of $\mathbf{S}^{-1}$ in (4.2) have interesting interpretations in discriminant analysis and logistic regression. By (3.10), the elements in the upper right-hand corner are the negatives of the discriminant coefficient estimates $\hat{\beta}_{jk}$. This might have been deduced from (4.2) and (3.18) by recognizing that $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}$ is the matrix of least-squares regression coefficients of the components of $\mathbf{x}$ on $v_1, \ldots, v_{m-1}$, and $\mathbf{S}_{11.2}$ is the residual sum of squares and cross-products matrix (Anderson, 1958, p. 81). Since the relevant regression equations are of the form (3.5) except that the term $\boldsymbol{\theta}_k'\mathbf{x}$ is missing, it follows that the columns of $\mathbf{S}_{12}\mathbf{S}_{22}^{-1}$ are the vectors $\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_m$, $j = 1, 2, \ldots, m - 1$, and $\mathbf{S}_{11.2} = \hat{\boldsymbol{\Sigma}}$.

Hence, if $\hat{\underset{\sim}{B}}$ is defined to be the q×(m - 1) matrix having $\hat{\underset{\sim}{\beta}}_j$ as its jth column, then it follows that

$$\underset{\sim}{S}^{-1} = \begin{pmatrix} \hat{\underset{\sim}{\Sigma}}^{-1} & - \hat{\underset{\sim}{B}} \\ - \hat{\underset{\sim}{B}} & \underset{\sim}{S}^{22} \end{pmatrix} . \tag{4.4}$$

By (3.9) and (3.13), the diagonal elements of $\underset{\sim}{S}^{22}$ are the multiples $K_j = n/SS_j$ used in converting the ILS estimates to the MLEs.

Clearly, the estimates $\hat{\beta}_{jk}$ and their test statistics can be calculated directly from $\underset{\sim}{S}^{-1}$. Following the linear model paradigm is simply a mnemonic technique for adopting standard least-squares procedures to isolate the appropriate elements of $\underset{\sim}{S}^{-1}$.

5. EFFICIENCY AND ROBUSTNESS

Logistic regression is often applied in situations in which the normality assumptions are known to be violated, e.g., in cases in which one or more of the independent variables are dichotomous. Several authors (e.g., Press and Wilson, 1978) have recommended against the use of the discriminant function estimators $\hat{\alpha}_j$ and $\hat{\underset{\sim}{\beta}}_j$ except in those rare instances when the normality assumptions apply.

A commonly recommended alternative to the discriminant function estimators when the normality assumptions do not apply are the conditional maximum likelihood estimators (CMLEs). These estimators are defined as the values of α_j and $\underset{\sim}{\beta}_j$ that maximize the conditional likelihood function

$$L_c = \prod_{j=1}^{m} \prod_{i=1}^{n_j} [p(j|\underset{\sim}{x}_i)]^{v_{ij}} , \qquad (5.1)$$

where $p(j|\underset{\sim}{x}) = P(y = j|\underset{\sim}{x})$ is given by (3.1) in the polytomous case and (1.3) in the dichotomous case. Of course, the CMLEs are the maximum likelihood estimators if the $\underset{\sim}{x}_i$'s are constant vectors or if the marginal distributions of the $\underset{\sim}{x}_i$'s do not depend on α_j and $\underset{\sim}{\beta}_j$, but even in these cases the rationale for adopting the CMLEs in practice is unclear.

Under the normality assumptions imposed in the preceding sections, one would expect that the discriminant estimators, being the <u>unconditional</u> MLEs, would perform at least as well as the CMLEs in large samples. Efron (1975) confirmed this in the dichotomous case and showed that the asymptotic efficiency of the CMLEs decreases markedly as the Mahalanobis distance between the mean vectors $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$ increases.

Despite this lack of efficiency in the normal case, it is often contended that the CMLEs are preferable because they are more robust in the nonnormal case. In any case, the CMLEs raise thorny computational and theoretical problems, and there may be some difficulty in determining whether the CMLEs exist for a given sample in the polytomous case. In the dichotomous case, the CMLEs do not exist if there is some linear combination $\underset{\sim}{d}'\underset{\sim}{x}$ such that the values $\underset{\sim}{d}'\underset{\sim}{x}_i$ for those individuals having $y_i = 1$ are all larger (or smaller) than the corresponding values of those individuals for which $y_i = 0$. If the CMLEs exist, they are often calculated using an iterative procedure, such as the method introduced by Walker and Duncan (1967), that may require a number of passes through the data. Test statistics associated with the CMLEs are based on asymptotic properties of MLEs that are of questionable validity in small samples.

While the t-statistics associated with the discriminant function estimators provide exact tests when the normality assumptions apply, the robustness of these statistics is open to question when the normality assumptions are violated. To provide some evidence on this score, consider the case in which there is just a single independent variable x in the dichotomous logistic model. It follows from the identification of the t-statistics (3.14) and (3.15) that the statistic $t = \hat{\beta}/s.e.(\hat{\beta})$ associated with the coefficient β on x is the ordinary two-sample t-statistic

$$t = (\bar{x}_1 - \bar{x}_2)/[S^2(n_1^{-1} + n_2^{-1})]^{1/2} \quad . \qquad (5.2)$$

As is well known, two-sided tests and confidence intervals based on this statistic are quite robust to departures from normality, even when the x_i's are dichotomous.

A case that would seem to favor the CMLEs is the case in which the $\underset{\sim}{x}_i$'s are constant vectors. Berkson (1955) made a thorough study of the performance of the CMLEs (here, MLEs) in bio-assay situations where the x_i's are preassigned dosage levels. He showed that his minimum logit chi-square estimators and the minimum chi-square estimators perform considerably better than the MLEs in applications of this type, even if one excludes the cases in which the MLEs fail to exist. The poor performance of the CMLEs in this case as well as the normal case raises questions about the widespread use of the CMLEs in practice.

This is not to say that the discriminant function estimators would perform any better in bio-assay situations. Halperin, Blackwelder, and Verter (1971) provide compelling arguments to effectively eliminate the discriminant function estimators as contenders in applications in which the independent variables are dichotomous. However, in cases like these, the data lend themselves to grouping, so that one can use the readily calculated minimum logit chi-square estimators. The procedure involves first transforming the group means using Berkson's logit transformation or a modified version recommended by Anscombe (1956) and then fitting the transformed values using weighted least squares. For an excellent discussion of these methods, see Cox (1970).

In applications where some or all of the independent variables are continuous, the discriminant function estimators merit wider use both in exploratory work associated with fitting logistic regression models and as alternatives to (as well as first approximations for) the CMLEs. The main reason for these recommendations stems from an empirical observation--the two methods ordinarily yield comparable results in practice. Both sets of estimates, their standard errors, and their t-statistics have been calculated for numerous data sets emanating from research studies at The Rand Corporation since 1974 when the formulas (2.3) for the dichotomous case were first derived (Haggstrom, 1974). Almost without exception, the results from applying the two procedures have been interchangeable for most practical purposes in that corresponding pairs of estimates typically differ by less than a standard error (no matter which standard error is used), and the t-statistics for the two procedures are usually quite close.

In their excellent paper comparing the two estimation techniques, Halperin et al. reported the results of using both procedures in fitting several data sets that included both continuous and discrete independent variables. For the most part, their results confirmed the close agreement of the estimation procedures, although they reported slightly better fits using the CMLEs. They found that the absolute values of the t-statistics associated with the CMLEs tended to be slightly smaller than those for the discriminant function estimates, but this may have resulted from their using a different t-statistic from the one defined in (2.21).

The observation that the two estimation procedures tend to yield comparable results (even in cases where the appropriateness of the logistic regression model is suspect) indicates that, whatever robustness properties the estimators have to nonnormality and misspecifications of the regression functions, the procedures seem to share those properties in situations where the CMLEs exist and some of the independent variables are continuous. Since neither procedure has been shown to have a decided advantage based on theoretical grounds (except perhaps in the normal case), it seems only reasonable to opt for the computational facility of the discriminant function estimators, especially in exploratory work with large data sets.

Another consideration that favors the use of the discriminant function estimators in some applications is that, unlike the CMLEs, they are readily adapted to handling missing values. As was seen in Section 4, the discriminant function estimates can be calculated directly from the augmented sample covariance matrix. In the missing values case, one can mimic a common procedure for handling missing values in linear models by simply substituting estimates of the elements s_{ij} using observations on complete pairs. Alternative procedures and software for carrying out this process is provided in BMDP-79 (Dixon and Brown, 1979, Chapter 12). Chow (1979) discusses this and other techniques for treating the missing value problem in logistic regression.

6. A NUMERICAL EXAMPLE

As an example to illustrate how a polytomous logistic regression model can be fitted by ordinary least squares, we report an analysis of 300 observations on participants in the National Longitudinal Study of the High School Class of 1972. The scores x_1 and x_2 are the seniors' Scholastic Aptitude Test scores (verbal and quantitative) divided by 100, and v_1, v_2, and v_3 are indicator variables for three categories of postsecondary activities: (1) College attendance, (2) military service, and (3) other. Some summary statistics for comparing the three groups are given in Table 1.

Table 1

SUMMARY STATISTICS

		Means		Std. dvn.		
Groups	n	x_1	x_2	x_1	x_2	$r(x_1,x_2)$
1	169	4.79	5.29	1.14	1.17	0.69
2	20	4.30	4.61	0.94	0.96	0.52
3	111	4.20	4.69	0.96	1.07	0.56
Combined	300	4.54	5.02	1.10	1.15	0.66

The equations below were fitted to the observations by ordinary least squares:

$$v_1 = -\ 1.0023 + \underset{(2.23)}{.0719}x_1 + \underset{(1.79)}{.0551}x_2 - \underset{(-5.27)}{.5610}v_2$$

$$v_2 = \quad .1525 + \underset{(0.77)}{.0131}x_1 - \underset{(-0.73)}{.0118}x_2 - \underset{(-5.27)}{.1531}v_1 \ .$$

The quantities in parentheses beneath the ILS estimates are the t-statistics $t = b_{jk}/s.e.(b_{jk})$. The multipliers for transforming the ILS estimates to the MLEs of the logistic regression coefficients (3.2) are $K_1 = n/SS_1 = 300/61.96 = 4.842$ and $K_2 = n/SS_2 = 300/16.91 = 17.74$. The values of the discriminant function estimates determined from (3.13) are reported in Table 2, along with the corresponding values of the CMLEs calculated from the same data set.

Table 2

ESTIMATES OF THE LOGISTIC REGRESSION COEFFICIENTS

	Discriminant function estimates			Cond. maximum likelihood		
	Coeff.	s.e.	t	Coeff.	s.e.	t
Group 1 (College)						
Constant	-2.476			-2.491		
x_1	.348	.156	2.23	.352	.157	2.24
x_2	.267	.149	1.79	.268	.148	1.81
Group 2 (Military)						
Constant	-1.731			-1.747		
x_1	.232	.300	.77	.235	.302	.78
x_2	-.209	.286	-.73	-.207	.286	-.72

In this particular case, the agreement between the discriminant function estimates and the CMLEs was remarkably close. While good agreement between the two sets of estimates and their t-statistics is expected, this level of agreement is unusual.

To illustrate that the discriminant function estimates and their t-statistics can be determined directly from the augmented sample covariance matrix $\underset{\sim}{S}$ for x_1, x_2, v_1, and v_2, the matrix and its inverse are given below:

$$\underset{\sim}{S} = \begin{bmatrix} 1.2029 & .8388 & .1415 & -.0158 \\ .8388 & 1.3298 & .1489 & -.0275 \\ -.1415 & .1489 & .2460 & -.0376 \\ -.0158 & -.0275 & -.0376 & .0622 \end{bmatrix}$$

$$\underset{\sim}{S}^{-1} = \begin{bmatrix} 1.5093 & -.9179 & -.3482 & -.2324 \\ .9179 & 1.3652 & -.2665 & .2089 \\ -.3482 & -.2665 & 4.8417 & 2.7168 \\ -.2324 & .2089 & 2.7168 & 17.7453 \end{bmatrix}$$

The negatives of the discriminant function estimates appear in the upper right-hand corner of $\underset{\sim}{S}^{-1}$, and the values of $K_j = n/SS_j$ are the last two diagonal elements. The t-statistics for the discriminant function estimates can be determined by first calculating the partial correlation coefficients $r_{jk.c} = - s^{jk}/[s^{jj}s^{kk}]^{1/2}$ and then applying formula (3.16).

The t-statistics in Table 2 associated with $\hat{\beta}_{21}$ and $\hat{\beta}_{22}$ can be used to test the hypotheses that $\beta_{21} = 0$ and $\beta_{22} = 0$. Alternatively, Hotelling's T^2 could be used to test the hypothesis that $\underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_3$. Given that these hypotheses are accepted, one might choose to combine Groups 2 and 3, thereby reducing the polytomous case to the dichotomous case. The requisite calculations for fitting the dichotomous model can be performed directly by inverting the appropriate 3x3 submatrix of $\underset{\sim}{S}$.

REFERENCES

Anderson, T. W., Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York, 1958.

Anscombe, F. J., "On Estimating Binomial Response Relations," Biometrika, Vol. 43 (1956), pp. 461-464.

Berkson, Joseph, "Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function," Journal of the American Statistical Association, Vol. 50 (1955), pp. 130-162.

Chow, Winston K., "A Look at Various Estimators in Logistic Models in the Presence of Missing Values," The American Statistical Association 1979 Proceedings of the Business and Economic Statistics Section, pp. 417-420.

Cox, D. R., Analysis of Binary Data, Chapman and Hall, Ltd., London, 1970.

Das Gupta, S., "Some Aspects of Discrimination Function Coefficients," Sankhyā, Series A, Vol. 30 (1968), pp. 387-400.

Day, N. E., and D. F. Kerridge, "A General Maximum Likelihood Discriminant," Biometrics, Vol. 23 (1967), pp. 313-323.

Dixon, W. J., and M. B. Brown (eds.), BMDP-79: Biomedical Computer Programs, P-Series, University of California Press, Berkeley, 1979.

Efron, Bradley, "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," Journal of the American Statistical Association, Vol. 70 (1975), pp. 892-898.

Ferguson, T. S., Mathematical Statistics: A Decision Theoretic Approach, John Wiley & Sons, Inc., New York, 1967.

Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. 7 (1936), pp. 179-188.

Fisher, R. A., "The Statistical Utilization of Multiple Measurements," Annals of Eugenics, Vol. 8 (1938), pp. 376-386.

Giri, N., "On the Likelihood Ratio Test of a Normal Multivariate Testing Problem," Annals of Mathematical Statistics, Vol. 35 (1964), pp. 181-190.

Haggstrom, G. W., "Notes on Logistic Regression and Discriminant Analysis," The Rand Corporation, 1974, unpublished.

Halperin, Max, W. C. Blackwelder, and J. I. Verter, "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," Journal of Chronic Diseases, Vol. 24 (1971), pp. 125-158.

Lachenbruch, P. A., Discriminant Analysis, Hafner Press, New York, 1975.

Lehmann, E. L., Testing Statistical Hypotheses, John Wiley & Sons, Inc., New York, 1959.

Press, S. J., and Sandra Wilson, "Choosing Between Logistic Regression and Discriminant Analysis," Journal of the American Statistical Association, Vol. 73 (1978), pp. 699-705.

Rao, C. R., "Tests with Discriminant Functions in Multivariate Analysis," Sankhyā, Vol. 7 (1946), pp. 407-414.

Rao, C. R., "Tests of Significance in Multivariate Analysis," Biometrika, Vol. 35 (1948), pp. 58-87.

Rao, C. R., Linear Statistical Inference and Its Applications, John Wiley & Sons, Inc., New York, 1965.

Walker, Strother H., and David B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," Biometrika, Vol. 54 (1967), pp. 167-169.

Warner, Stanley L., Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice, Northwestern University Press, Evanston, Illinois, 1962.